تولیدات دامی
دوره ۱۹ ■ شماره ۱ ■ بهار ۱۳۹۶
صفحه‌های ۱۲-۱

مقایسه روش‌های آماری پارامتری و بازنمونه‌گیری در ارزیابی صفات کمّی با ساختار ژنتیکی متفاوت

منوچهر مرادی[۱]، رستم عبداللهی آرپناهی[۲]*، بهزاد همتی[۳]، ابوالقاسم لواف[۳]

۱. دانش آموخته کارشناسی ارشد، دانشگاه آزاد اسلامی واحد کرج، کرج- ایران
۲. استادیار، گروه علوم دام و طیور، پردیس ابوریحان، دانشگاه تهران، تهران- ایران
۳. دانشیار، گروه علوم دامی، دانشکده کشاورزی و منابع طبیعی، دانشگاه آزاد اسلامی واحد کرج، کرج- ایران

تاریخ وصول مقاله: ۱۳۹۵/۰۳/۰۶ تاریخ پذیرش مقاله: ۱۳۹۵/۰۶/۳۰

چکیده

هدف از این مطالعه مقایسه سه روش پارامتری (GBLUP، BayesB، RKHS) و دو روش بازنمونه‌گیری (Bagging GBLUP و Random Forest) در پیش بینی ارزش‌های اصلاحی ژنومیک برای صفاتی با ساختار ژنتیکی متفاوت بود. یک ژنوم با سه کروموزوم، هر کروموزوم به طول یک مورگان شبیه‌سازی شد و روی آن ۱۵۰۰ نشانگر تک نوکلئوتیدی (SNP) در سه سناریو ۵۰، ۱۰۰ و QTL۲۰۰ به طور یکنواخت پخش شدند. اثر جایگزینی QTLها با استفاده از توزیع نرمال استاندارد، گاما و یکنواخت با وراثت‌پذیری ۳۰ درصد مدل سازی شدند. توانایی پیش‌بینی روش‌های آماری با استفاده از آماره‌های همبستگی بین ارزش‌های اصلاحی پیش‌بینی شده و واقعی و همچنین رگرسیون ارزش اصلاحی واقعی بر پیش‌بینی شده بررسی شد. نتایج نشان داد در جمعیت‌های تایید، روش RF باعث بیش-برآورد رگرسیون ارزش‌های اصلاحی واقعی بر پیش‌بینی شده شد، در حالی که روش‌های GBLUP، BayesB و RKHS منجر به کم-برآورد ضریب رگرسیون شدند. به جز روش Bagging GBLUP در دیگر روش‌ها تفاوت معنی داری با تغییر توزیع اثرات QTL مشاهده نشد اما در مجموع عملکرد دو روش GBLUP و BayesB نسبت به دیگر روش‌ها بهتر بود. یکی از دلایل احتمالی برتری GBLUP و BayesB بر دیگر روش‌ها می‌تواند شبیه سازی صفات با اثرات صرفاً ژنتیکی افزایشی بوده باشد. به طور کلی، روش‌های GBLUP و BayesB بر روش های بازنمونه‌گیری در پیش‌بینی های ژنومی ارجحیت دارند.

کلیدواژه‌ها: ارزیابی ژنومی، جنگل تصادفی، فضای تولید هسته هیلبرت، کیسه بندی شده، معماری ژنتیکی، یادگیری ماشین

* نویسنده مسئول Email: r.abdollahi@ut.ac.ir

مقدمه

با فراهم شدن تراشه‌های SNP با تعداد ۸۰۰-۳ کیلو باز، روش انتخاب به کمک کل ژنوم یا انتخاب ژنومی جایگزین روش‌های انتخاب کلاسیک براساس روش بهترین پیش‌بینی نااریب خطّی (BLUP) و انتخاب به کمک نشانگرها (MAS) شده است. با انتخاب ژنومی فاصله نسل کم شده و صحت انتخاب زیاد می‌شود و در نتیجه پیشرفت ژنتیکی افزایش می‌یابد [۲۳].

عواملی نظیر اندازه جمعیت مرجع، تعداد نشانگر SNP، مقدار عدم تعادل پیوستگی، وراثت‌پذیری صفت، معماری ژنتیکی صفت و روش آماری مورد استفاده بر صحت ارزیابی‌های ژنومی تاثیر دارند. روش‌های مختلفی نظیر حداقل مربعات، بهترین پیش‌بینی نااریب خطی ژنومی و انواع روش‌های بیزی برای برآورد اثر نشانگرها پیشنهاد شده است [۸]. دقت پیش‌بینی هر کدام از این روش‌ها نسبت به دیگری متفاوت است [۷، ۸]. بیشتر روش‌های برآورد اثر نشانگرها، مدل‌های خطی هستند که اثر SNPها از طریق رگرسیون فنوتیپ یا ارزش اصلاحی بر نشانگرهای SNP برآورد می‌شود [۹]. اخیراً روش‌های یادگیری ماشین نیز در انتخاب ژنومی مطرح شده‌است [۱۱، ۱۸]. یکی از مزایای روش‌های یادگیری ماشین توانایی آنها در تجزیه و تحلیل داده‌های با تعداد بسیار زیاد است [۱۲]. در آینده‌ی نزدیک و با در دسترس بودن اطلاعات ژنوتیپی یا اطلاعات توالی ژنوم با تعداد بسیار زیاد، این روش‌ها به خوبی برای تجزیه و تحلیل این داده‌ها کارایی خواهند داشت [۱۴]. همچنین، بررسی روابط پیچیده بین متغیرها (مانند اثرات متقابل بین نشانگرها) نیز از دیگر مزایای این روش‌ها است [۱۳]. کاربرد این روش‌ها در پیش‌بینی ارزش‌های اصلاحی ژنومی بسیار جدید است و به دلیل ویژگی‌های مطلوب آنها استفاده از آنها در حال گسترش است [۱۱].

اگرچه، توانایی تعدادی از روش‌های بازنمونه گیری نظیر GBLUP کیسه بندی شده (Bagging GBLUP)، جنگل تصادفی (RF) [۱۰، ۱] و روش‌های ناپارامتری نظیر شبکه عصبی (NN)، ماشین بردار پشتیبانی (SVM) [۱۲، ۱۳ و ۲۱] در ارزیابی ژنومی بررسی شده‌اند، ولی مقایسه جامعی بین آنها انجام نشده است. هدف از این پژوهش مقایسه روش نیمه پارامتری فضای تولید هسته هیلبرت (RKHS) و دو روش یادگیری شورایی Bagging GBLUP و RF با روش‌های آماری استاندارد ارزیابی ژنومی نظیر GBLUP و BayesB در پیش‌بینی ارزش‌های اصلاحی ژنومیک تحت تاثیر تعدادی از اجزای معماری ژنتیکی صفات کمّی مانند توزیع اثرات ژنی و تعداد QTL بود.

مواد و روش‌ها

با استفاده از بسته نرم افزاری hypred [۲۴]، یک ژنوم با سه کروموزوم هر کدام به طول یک مورگان شبیه سازی شد. روی هر کروموزم ۵۰۰ نشانگر تک نوکلئوتیدی (SNP) دو آللی با فراوانی اولیه یکسان ۵۰ درصد در سه سناریو QTL معادل ۵۰، ۱۰۰ و ۲۰۰ به طور یکنواخت توزیع شدند. اثر جایگزینی QTLها با استفاده از توزیع نرمال استاندارد، گاما و یکنواخت با تغییرات لازم در کدهای برنامه Hypred و با وراثت‌پذیری ۳۰ درصد مدل‌سازی شد (جدول۱). برای هر جایگاه SNP با ژنوتیپAA کد دو، با ژنوتیپAa کد یک و با ژنوتیپaa کد صفر منظور شد. جمعیت پایه به تعداد ۱۰۰ حیوان (۵۰ نر و ۵۰ ماده) شبیه‌سازی شد و برای ۵۰۰ نسل آمیزش تصادفی انجام شد. در این حالت به طور تصادفی از هاپلوتایپ‌های پدری و مادری نمونه‌گیری و از آنها برای تولید نتاج استفاده شد. از هر دو والد فقط دو فرزند ایجاد شد، در نتیجه اندازه جمعیت در طی ۵۰۰ نسل ثابت بود. تحت شرایط ثابت بودن اندازه جمعیت، رانش ژنتیکی و نوترکیبی منجر به ایجاد تعادل موتاسیون- رانش ژنتیکی

شد و در این حالت بین نشانگرها و QTLها پیوستگی ایجاد شد. در نسل ۵۰۱ اندازه جمعیت به ۱۰۰۰ حیوان افزایش داده شد. از این حیوانات با اطلاعات ژنوتیپی، فنوتیپی و همچنین ارزش‌های اصلاحی ژنومی جمعیت مرجع تشکیل شد. سپس، نسل‌های ۵۰۲ تا ۵۰۴ از افراد نسل ۵۰۱ ایجاد شدند که این حیوانات دارای اطلاعات ژنوتیپی بودند ولی اطلاعات فنوتیپی نداشتند. در واقع این نسل‌ها جمعیت‌های تأیید بودند که ارزش‌های اصلاحی ژنومی آنها پیش‌بینی شد. برنامه شبیه سازی پنج بار تکرار شد و میانگین و انحراف معیار نتایج برای انجام مقایسات لازم ذخیره شد.

جدول ۱. پارامترهای استفاده شده در شبیه‌سازی جمعیت

ژنوم	وضعیت
اندازه ژنوم	۳ مورگان
تعداد کروموزوم	۳
تعداد نشانگر SNP در هر کروموزوم	۵۰۰
تعداد QTL در هر کروموزوم	۵۰ ، ۱۰۰ و ۲۰۰
توزیع اثرات QTL	یکنواخت، نرمال، گاما
جمعیت	
تعداد نسل‌ها برای ایجاد عم تعادل پیوستگی (LD)	۱-۵۰۰
نسل مرجع	۵۰۱
نسل‌های تایید	۵۰۲-۵۰۴
اندازه موثر جمعیت تاریخی (*Ne*)	۱۰۰

ارزیابی‌های ژنومی با پنج روش GBLUP، BayesB [۲، ۸]، Bagging GBLUP، Random Forest و RKHS انجام شد.

کلمه Bagging مخفف Bootstrap aggregating sampling است و یک روش بازنمونه گیری یا شورایی است. ثابت شده است وقتی واریانس پیش‌بینی متغیرهای مستقل یا پیش‌بینی کننده‌ها زیاد باشد با استفاده از این روش صحت پیش‌بینی‌ها افزایش می‌یابد [۵، ۲۵]. هنگامی که کم و زیاد کردن متغیرهای مستقل باعث تغییرات زیاد در جواب‌ها می‌شود کارایی روش Bagging مناسب است [۱، ۵]. معمولاً این حالت در مواردی است که همبستگی بین متغیرهای مستقل زیاد باشد [۱].

اگر یک مجموعه داده (D) به اندازه N مشاهده وجود داشته باشد، در روش Bagging به تعداد B مجموعه داده‌های جدید (D_b) از جمعیت اولیه با جایگزینی انتخاب می‌شود [b=1,2,…,B]. با نمونه گیری با جایگزینی ممکن است تعدادی از مشاهدات در هر مجموعه D_b چندین بار تکرار شوند و این به عنوان Bootstrap Sampling شناخته شده است. یک مدل مشابه برای هر یک از B نمونه bootstrap استفاده شده و در پایان میانگینی از خروجی‌ها محاسبه می‌شود. روش Bagging اولین بار در قالبGBLUP ارایه شد [۱۰] که GBLUP کیسه بندی شده (Bagging GBLUP) نامیده می‌شود. این روش هنگامی که پیش‌بینی کننده ها (نشانگرها) ناپایدار هستند و یا

همبستگی بین متغیرهای مستقل (مانند نشانگرهای SNP) زیاد است، به دلیل قابلیت کاهش واریانس خطا، پیش‌بینی دقیق‌تری ارائه می‌دهد [۱].

الگوریتم یادگیری جنگل تصادفی (RF) در قالب بسته نرم افزاری *randomForest* [۴] برای پیش‌بینی ارزش‌های اصلاحی ژنومی استفاده شد. روش RF شامل مجموعه‌ای از درختان رگرسیونی است که هر کدام با استفاده از اطلاعات ورودی n نمونه (شامل اطلاعات ژنوتیپی و فنوتیپی افراد جمعیت مرجع) ایجاد می‌شود. مدل در جمعیت مرجع آموزش داده می‌شود و در جمعیت تأیید (حیوانات کاندیدای انتخاب) اعمال می‌شود. یکی از n نمونه وارد هر گره از هر درخت می‌شود (*mtry*) و از این نمونه اطلاعات یک SNP برای تقسیم‌بندی حیوانات استفاده می‌شود، و حیوانات بر اساس اطلاعات ژنوتیپی دسته‌بندی می‌شوند. این کار در گره‌های متوالی انجام می‌شود و در پایان به برگ‌ها و یا همان گره‌های پایانی ختم می‌شود که در آنها حداکثر یکنواختی وجود دارد (حیوانات دارای اطلاعات فنوتیپی با ژنوتیپ‌های مشابه برای SNPهای مختلف در گره پایانی تجمع می‌یابند). پیش‌بینی RF برای یک مثال ورودی جدید (حیوان دارای اطلاعات ژنوتیپی x_iاما فاقد اطلاعات فنوتیپی y_i)، $\hat{f}_{rf}^{B}(x)$، از طریق محاسبه میانگین از B درخت، $[T(x,\Psi_b)]_1^B$، و به کمک رابطه ۱ محاسبه شد.

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x,\Psi_b), \qquad \text{رابطه (۱)}$$

در این رابطه ، $\boldsymbol{\Psi_b}$، bامین درخت در RF را نشان می‌دهد. پارامترهای مهم درRF تعداد متغیرهای انتخاب شده در هر گره درخت (mtry)، تعداد درخت (ntree) و حداقل اندازه یا حداقل تعداد مشاهده‌ها در گره‌های پایانی یا برگ‌ها می‌باشند که قبل از انجام آنالیزها باید مقدار مناسب آنها تعیین شود. در ارتباط با داده‌های پیوسته مقدار پیشنهاد شده برای تعداد متغیر انتخاب شده در هر گره درخت برابر p/3 است که p تعداد SNP است. در هر بار نمونه‌گیری با جایگزینی از اطلاعات، تعدادی اطلاعات (SNPها) نمونه‌گیری نمی‌شوند و تعدادی دیگر شاید چند بار نمونه‌گیری شوند. به عبارت دیگر هر داده ورودی برای تعدادی درخت‌ها داده‌های خارج از کیسه (داده‌های نمونه گیری نشده) خواهند. این داده‌ها به عنوان یک معیارسنج داخلی برای هر درخت عمل می‌کنند و اعتبارسنجی از طریق برآورد خطای خارج از کیسه (OOB error) انجام می‌شود. اگر داده‌های خارج از کیسه از طریق درختان پیش‌بینی شوند، برای این پیش‌بینی‌ها خطا وجود دارد و میانگین آنها را خطای خارج از کیسه می‌نامند که نشان دهنده میزان تأثیر نمونه‌های انتخاب نشده بر میزان خطای نتیجه نهایی RF است [۴].

استفاده از روش RKHS در سال ۲۰۰۶ با کاربرد در مدل‌های مختلط ارائه شد [۹]. رگرسیون هسته‌ای ریج بیزین شکلی از روش‌های ناپارامتری فضای تولید هسته‌ای هیلبرت (RKHS) است [۲۰]. فرض شود که ارتباط فنوتیپ باSNP برای حیوان i از مدل ۲ پیروی می‌کند.

$$y_i^R = \mu + g(x_i) + \epsilon_i \qquad \text{رابطه (۲)}$$

در این رابطه، $g(x_i)$، بردار ژنوتیپ‌های SNP حیوان iام است . فرض می‌شود که g به صورت Kα می‌باشد و K یک ماتریس هسته‌ایn×n بوده ارتباط بین ژنوتیپ‌های SNP در افراد مختلف را نشان می‌دهد. کاهش ابعاد از تعداد نشانگر (p) به تعداد حیوان (n) باعث می‌شود که بتوان بر تعداد متغیّرهای مجهول غالب شد و تعداد معادله را به اندازه n کاهش داد. در این حالت g= Kα، ابعاد ماتریس ژنومی از تعداد SNP به تعداد مشاهدات کاهش می‌یابد. اگر مجموع مربعات باقی مانده و نرم ضرایب α به ترتیب به عنوان تابع زیان و جریمه در نظر گرفته شوند این بیزین ریج رگرسیون بوده که هسته رگرسیون ریج K جایگزین ماتریس ژنوتیپی (X) با ابعاد n×p شده است.

اکنون می توان معادله مزبور را در فرم ماتریسی (رابطه ۳) نوشت.

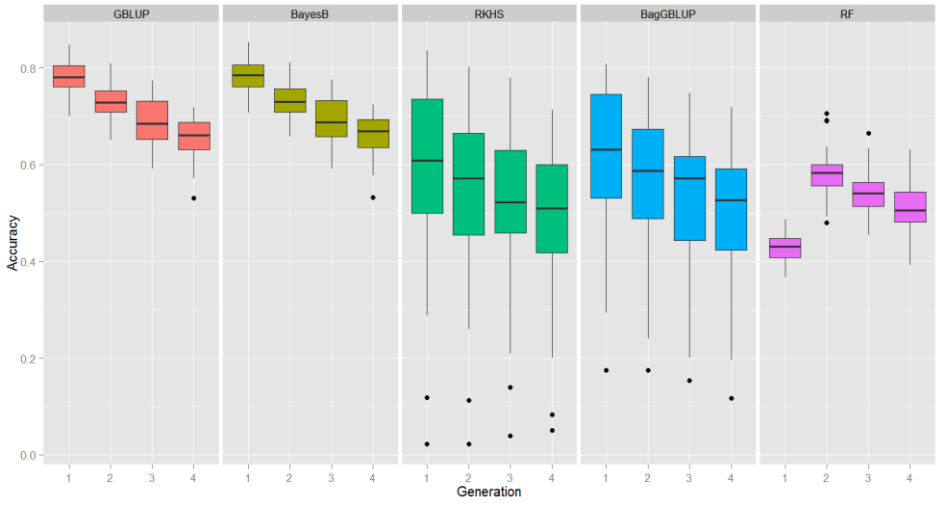
$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{K\alpha} + \boldsymbol{\varepsilon} \qquad \text{رابطه (۳)}$$

صحت و دقت ارزش‌های اصلاحی ژنومی پیش‌بینی شده توسط مدل‌های مختلف مقایسه شد. صحت و دقت پیش‌بینی از طریق همبستگی بین ارزش‌های اصلاحی ژنومی پیش‌بینی شده و ارزش‌های اصلاحی ژنومی واقعی (شبیه‌سازی شده) و رگرسیون ارزش‌های اصلاحی پیش‌بینی شده بر ارزش‌های اصلاحی واقعی برآورد شد. در مواردی که تفاوت بین روش‌ها زیاد بود با استفاده از آزمون t - استودنت نتایج مقایسه شد.

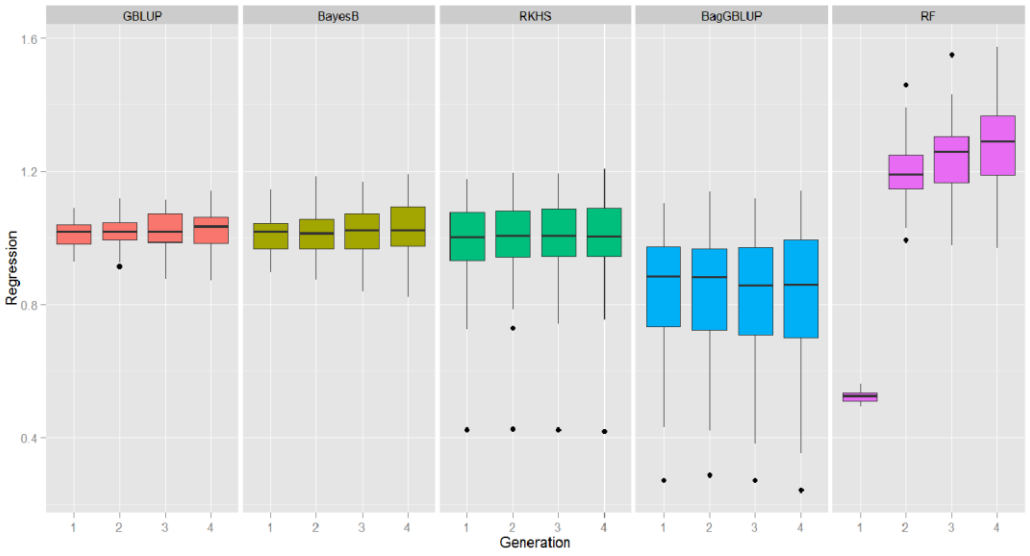
نتایج و بحث

با افزایش فاصله ژنتیکی بین جمعیت مرجع و جمعیت تأیید صحت پیش‌بینی ارزش‌های اصلاحی ژنومی کاهش یافت ($p<۰/۰۵$؛ شکل ۱). در یک پژوهش صحت پیش‌بینی ارزش‌های اصلاحی ژنومی با افزایش فاصله از جمعیت مرجع در طی پنج نسل از ۰/۸۴۸ (اولین نسل از حیوانات کاندیدای انتخاب) به ۰/۷۱۸ در نسل پنجم کاهش یافت [۲]. همچنین صحت پیش‌بینی ارزش‌های اصلاحی ژنومی برآورد شده با افزایش فاصله از جمعیت تأیید از ۰/۴۷ (در نسل اول) به ۰/۰۷ (در نسل ۱۰) کاهش یافت [۳]. در ضمن در نسل‌های آخر نسبت به نسل‌های اول اشتباه استاندارد (SD) صحت پیش‌بینی، افزایش یافت. به عبارت دیگر کارایی روش‌های برآورد اثرات نشانگری و تکرارپذیری نتایج کاهش یافته است.

نشانگرها در دو حالت تعادل و یا عدم تعادل با QTLها هستند. اگر نشانگرها در حالت تعادل با QTLها باشند، میزان کاهش در صحت پیش‌بینی با افزایش فاصله از جمعیت مرجع زیاد است. اگر نشانگرها در حالت عدم تعادل با QTLها باشند با افزایش فاصله از جمعیت تأیید، الگوی تعادل بین نشانگرها و QTLها نسبت به جمعیت مرجع تغییر می‌کند. بنابراین نمی‌توان از نشانگرها به خوبی اثرات QTLها را شناسایی نمود. ولی سرعت کاهش صحت پیش‌بینی ناشی از افزایش فاصله از جمعیت مرجع کمتر از حالتی است که نشانگرها در حالت تعادل پیوستگی با QTLها باشند [۲۱، ۲۶]. واریانس برآورد شده حاصل از روش‌های RKHS و Bagging GBLUP زیاد و در روش GBLUP و BayesB کم است.



شکل ۱. مقایسه صحت ارزش‌های اصلاحی ژنومی در طی چهار نسل و پنج روش GBLUP، BayesB، RKHS، BaggGBLUP و Randm Forest

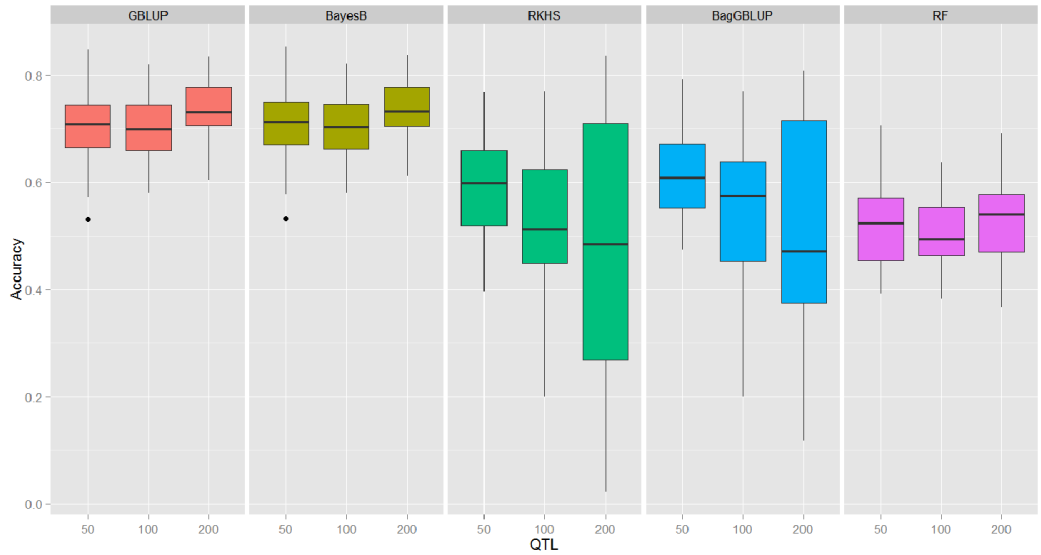


شکل ۲. مقایسه رگرسیون ارزش‌های اصلاحی ژنومی واقعی بر پیش‌بینی شده در طی چهار نسل و پنج روش GBLUP، BayesB، RKHS، BaggGBLUP و Randm Forest

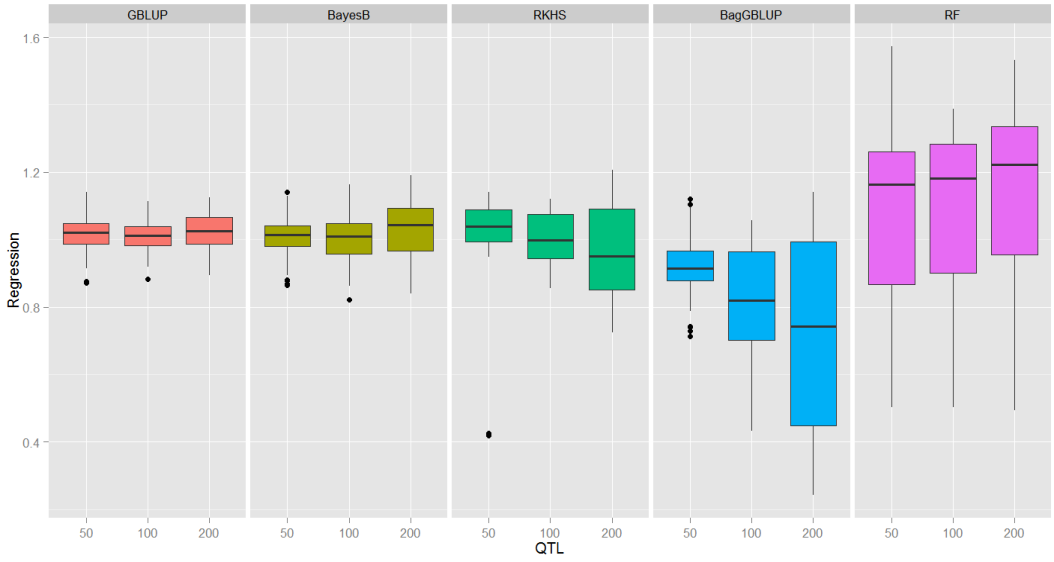
نتایج حاصل از رگرسیون ارزش‌های اصلاحی ژنومی واقعی از پیش‌بینی شده در طی چهار نسل (نسل اول مرجع و نسل دو، سه و چهار تایید) در شکل ۲ ارائه شده است. برآوردهای حاصل از روش‌های پارامتری و نیمه پارامتری (GBLUP، BayesB و RKHS) نزدیک به یک و ناریب بود ولی واریانس برآوردهای حاصل از روش‌های بازنمونه‌گیری نظیر RF، GBLUP Bagging زیاد است. در روش RF در جمعیت مرجع بیش- براورد ارزش‌های اصلاحی ژنومی حاصل شد ولی در نسل های تایید بیش‌برآورد ضریب رگرسیون و یا به عبارتی میزان اریبی ارزش‌های اصلاحی به طرف کم بود. رگرسیون ارزش‌های اصلاحی واقعی بر ارزش‌های اصلاحی پیش‌بینی شده یک معیار مهم است که باید در مطالعات ژنومی گزارش شود [۷]. در جمعیت‌های حیوانات اهلی، با گذشت زمان و زیادشدن فاصله از جمعیت مرجع مقدار LD به دلیل نوترکیبی، انتخاب و مهاجرت تغییر می‌کند [۱۵، ۱۹، ۲۲].

با تغییر تعداد QTL صحت پیش‌بینی دو روش GBLUP و BayesB مشابه و در روش RF کمتر بود. به طور کلی صحت پیش‌بینی حاصل از دو روش بیز B و GBLUP با تغییر تعداد QTL مشابه بود. در دو روش BagGBLUP و RKHS با افزایش تعداد QTL صحت ارزش‌های اصلاحی نوسان زیادی داشت. کاهش صحت پیش‌بینی ارزش‌های اصلاحی ژنومی در نتیجه افزایش تعداد QTL از ۵۰ به ۲۰۰ در مطالعات قبل گزارش شده است [۲]. در تحقیق مزبور تعداد ۵۰۰ SNP در نظر گرفته که کمتر از تعداد حیوانات بوده است ولی تحقیق حاضر تعداد SNP بیشتر از تعداد افراد جمعیت مرجع در نظر گرفته شده است [۲۰].

در حالتی که تعداد QTLها زیاد می‌شود، تعداد نشانگر مورد نیاز برای برآورد اثر کلیه QTLها بیشتر است [۷]. یعنی با زیاد شدن تعداد QTL در صورتی صحت پیش‌بینی ارزش‌های اصلاحی ژنومی افزایش می‌یابد که تعداد نشانگرها نیز افزایش یابد.



شکل ۳. مقایسه صحت ارزش‌های اصلاحی ژنومی در سه تعداد QTL (۵۰، ۱۰۰ و ۲۰۰) و پنج روش GBLUP، BayesB، RKHS، BaggGBLUP و Randm Forest



شکل ۴. مقایسه رگرسیون ارزش‌های اصلاحی ژنومی واقعی بر پیش‌بینی شده در سه تعداد متفاوت QTL (۵۰، ۱۰۰ و ۲۰۰) و پنج روش GBLUP، BayesB، RKHS، BaggGBLUP و Randm Forest

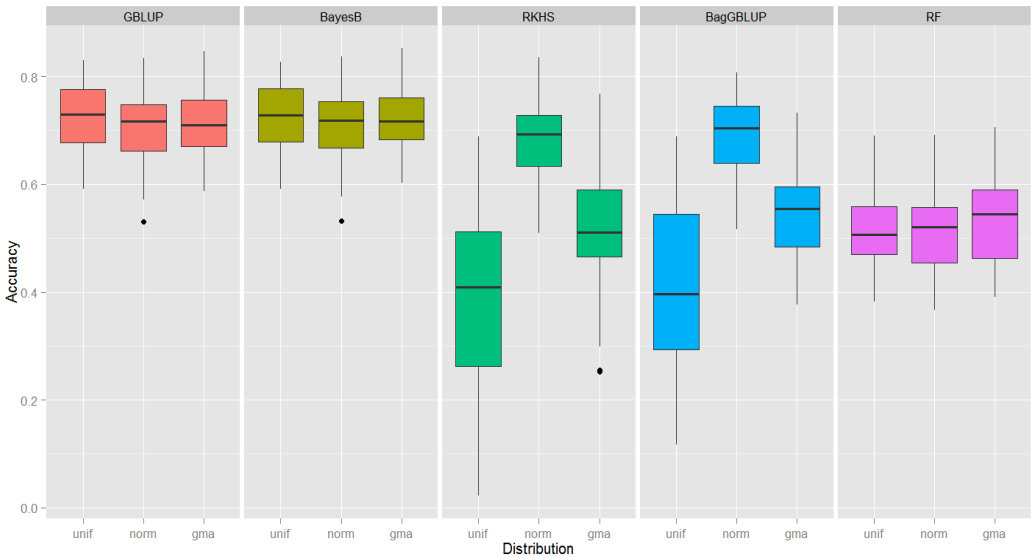
در حالتی که مقدار ضریب رگرسیون ارزش‌های اصلاحی واقعی بر پیش‌بینی شده بیشتر یا کمتر از یک باشد به ترتیب برآورد ارزش‌های اصلاحی کمتر و بیشتر از مقدار واقعی است. رگرسیون ارزش اصلاحی واقعی بر مشاهده شده در دو روش GBLUP و BayesB در هر سه تعداد QTL یکسان بود. واریانس برآورد در دو روش شورایی RF، GBLUP Bagging زیاد بود ولی در دو روش GBLUP Bagging در تعداد QTL زیاد، براوردها بیشتر و روش RF کمتر از مقدار واقعی بود.

تغییرات صحّت ارزش‌های اصلاحی ژنومی با تغییر

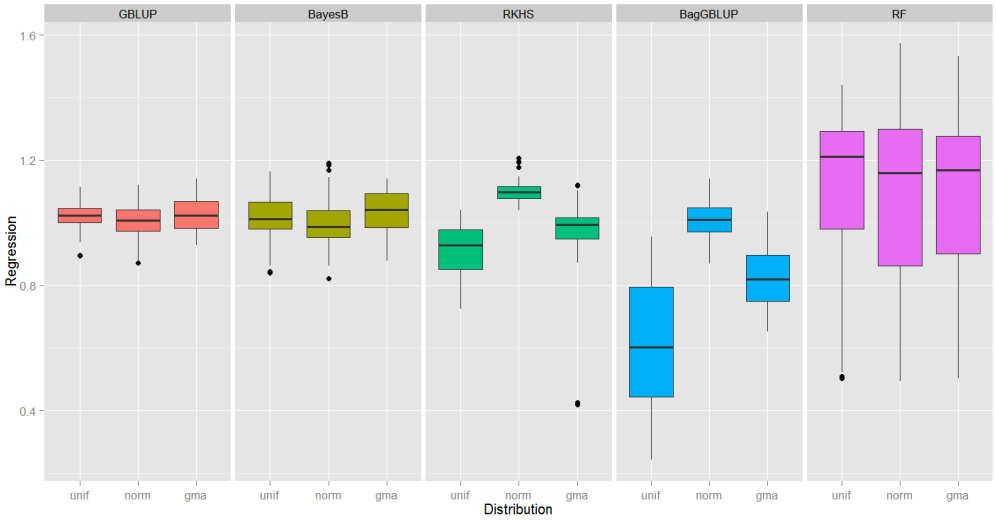
تعـداد QTL در دو روش اسـتاندارد GBLUP و BayesB مشـابه بـود. صـحت پـیش‌بینـی در دو روش GBLUP Bagging و RKHS در حالت نرمـال بـودن توزیـع اثـرات ژنی نسبت به دو توزیع گاما و یکنواخت بیشـتر بـود (P< 0.05). صحت روش RF در هر سه تعداد QTL نسبت بـه دیگر روش‌ها کمتر بود (شکل ۵).

ضرایب رگرسـیون در دو روش اسـتاندارد GBLUP و BayesB نااریب بود (شکل ۶). در حالتی که توزیع اثـرات ژنـی نرمـال بـود مقـدار ضـریب رگرسـیون در دو روش RKHS و Bagging GBLUP نزدیـک بـه یـک و نااریـب حاصل شد ولی در دو توزیع اثرات ژنی یکنواخت و کمتـر از مقدار واقعی برآورد شد. صحت ارزش‌های اصلاحی در دیگـر روش‌هـا (GBLUP, BayesB, RF)، در سـه توزیـع اثرات ژنی مشابه بود. مشابه سایر حالت‌های قبل، واریانس برآورد درروش RF زیاد بود.

در حالت توزیع اثـرات ژنـی گامـا از میـان روش‌هـای بیزین، روش بیزB و بیزL بیشـترین صـحت را دارنـد [۲]. درحالتی که توزیع اثرات ژنی گاما است تعـدادی از ژن‌هـا داری اثر عمده و درصد زیادی از ژن‌ها اثـرات نزدیـک بـه صفر دارند. بنابراین صـحت روش‌هـای جریمـه‌ای همـراه انتخاب متغیر مانند بیزB بهتر از دیگر روش‌هایی است، کـه فقط جریمه‌ای عمل می کنند. با استفاده از داده‌های واقعـی صفت تولیـد شـیر و نـرخ آبسـتنی در گاوهـای هلشـتاین فرانسوی، دلیل مشابه بودن نتایج حاصل از روش‌های بیزی و GBLUP در پیش بینی ارزش‌هـای اصـلاحی ژنـومی را معماری ژنتیکی پلی ژنی بودن و توزیع نرمال اثـرات ژنـی دانسته‌اند [۶]. روش‌های RRBLUP و GBLUP برای همه نشانگرها سهم یکسانی در پـیش بینـی ارزش‌اصـلاحی در نظر می‌گیرند ولی روش‌های بیزی بر حسب توزیع پیشین، وزن‌های متفاوتی به SNPها می‌دهند. در مطالعـات شـبیه-سازی قبل، روش‌های پارامتری [۱] و تعدادی از روش‌های پارامتری و غیرپـارامتری [۱۶] در انتخـاب ژنـومی بـا هـم مقایسه شده‌اند. در همـه ایـن تحقیقـات معمـاری ژنتیکـی صفت همانند این مطالعه به صورت افزایشی شـبیه سـازی شده بود و پیش‌بینی شد برای معماری ژنتیکـی اپیسـتازی، روش‌های شورایی یا ناپارامتری بهتر خواهند بود [۱۷].



شکل ۵. مقایسه صحت ارزش‌های اصلاحی ژنومی در سه توزیع اثرات جایگزینی QTL (یکنواخت=unif، نرمال=norm و گاما=gma) و پنج روش GBLUP، BayesB، RKHS، BaggGBLUP و Randm Forest



شکل ۶. مقایسه رگرسیون ارزش‌های اصلاحی ژنومی پیش‌بینی شده بر واقعی در سه توزیع اثرات جایگزینی QTL **(یکنواخت=unif، نرمال=norm و گاما=gma) و پنج روش** GBLUP، BayesB، RKHS، BaggGBLUP و Randm Forest

نتایج یک مطالعه شبیه‌سازی با مقایسه ۱۰روش پارامتری و چهار روش ناپارامتری برای صفاتی با معماری ژنتیکی کاملاً افزایشی و یا کاملاً اپیستازی نشان داد در حالتیکه معماری ژنتیکی به طور کامل بر پایه اپیستازی باشد، صحت روش‌های ناپارامتری بهتر از روش‌های پارامتری است. در حالی که برای معماری ژنتیکی کاملاً افزایشی صحت پیش‌بینی روش‌های پارامتری بهتر از روش‌های ناپارامتری است [۱۷].

به طور کلی نتایج این تحقیق نشان داد که صحت پیش‌بینی ارزش‌های اصلاحی در تعداد متغیر QTL با هم مشابه است و برآورد ارزش‌های اصلاحی با روش‌های شورایی (bagging GBLUP و RF) اریب است. در حالی که روش‌های پارامتری (BayesB و GBLUP) پیش‌بینی‌های نااریب حاصل شد. دو روش RKHS و Bagging GBLUP وقتی توزیع اثرات QTL نرمال بود نسبت به دو توزیع اثرات یکنواخت و گاما صحت بالاتری داشتند. در مجموع صحت پیش‌بینی روش‌های GBLUP و BayesB نسبت به دیگر روش‌های نیمه‌پارامتری و ناپارامتری بهتر بود.

منابع

۱. عبدالهی آرپناهی ر، پاکدل ع، نجاتی-جوارمی ا و مرادی شهربابک م (۱۳۹۲) مقایسه روش‌های ارزیابی ژنومیک در صفاتی با معماری ژنتیکی گوناگون. مجله تولیدات دامی، ۱۵(۱): ۷۷-۶۵.

2. Abdollahi-Arpanahi R, Morota G, Valente BD, Kranis A, Rosa GJM and Gianola D (2015) Assessment of bagging GBLUP for whole genome prediction of broiler chicken traits. Journal of Animal Breeding and Genetics. 132(3): 218-228.

3. Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAMand Bovenhuis H (2011) Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genetics Selection Evolution. 44:3.

4. Boulesteix AL, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Technical Report. Department of Statistics. University of Munich.

5. Breiman L. (1996) Bagging predictors. Machine Learning., 24, 123-140.

6. Colombani C, Legarra A, Fritz S, Guillaume F, Croiseau P, Ducrocq V and Robert-Granié C (2012) Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCp methods for genomic selection in French Holstein and Montbéliarde breeds. Journal of Dairy Science. 96: p. 575–591.

7. Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G and Hickey JM (2013) Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193: 347–365.

8. de los Campos G, Hickey JM. Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2), 327-345.

9. Gianola D, Fernando RL and Stella A (2006) Genomic-assisted prediction of genetic value with semi-parametric procedures. Genetics. 173:1761-1776.

10. Gianola D, Weigel KA, Krämer N, Stella A, Schön C-C (2014). Enhancing genome-enabled prediction by bagging genomic BLUP. *PLoS ONE*, 9, e91693.

11. González-Camacho JM, de Los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics. 125(4):759-71.

12. González-Recio O and Forni S (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution 43:7.

13. González-Recio O, Jiménez-Montero AJ and Alenda R (2013) The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. Journal of Dairy Science 96: 614–624.

14. González-Recio O, Rosa GJ, Gianola D (2014) Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livestock Science. 166:217-31.

15. Habier D, Fernando RL and Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182; 343–353.

16. Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: a comparison of models. Crop Science. 52:146-160.

17. Howard R, Carriquiry AL, Beavis WD (2014) Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. G3: Genes| Genomes| Genetics: g3. 114.010298.

18. Long N, Gianola D, Rosa GJM, Weigel KA and Avendano S (2007) Machine learning classification procedure for selecting SNPs in genomic selection: Application to early mortality in broilers. Journal of Animal Breeding and Genetics 124: 377–389.

19. Meuwissen THE, Hayes BJ and Goddard ME (2001) Prediction of total genetic value using genome wide dense marker maps. Genetics. 157: 1819–1829.

20. Morota G, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: a review. Frontiers in genetics. 5:363.

21. Moser G, Tier B, Crump RE, Khatkar MS and Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution. 41:56.

22. Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics. 124: 342-355.

23. Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. Journal of 3- Animal Breeding and Genetics 123: 218–223.

24. Technow FR (2013) hypred: Simulation of genomic data in applied genetics. Available at: http://cran.r-project.org/web/packages/hypred/index.html.

25. Valle C, Nanculef R, Allende H and Moraga C (2007). Two bagging algorithms with coupled learners to encourage diversity.In, *Advances in Intelligent Data Analysis VII*. Springer. pp. 130-139.

26. Wolc A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Habier D, Fernando R, Garrick DJand Dekkers JC (2011) Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genetics Selection Evolution. 43, 23.

Journal of
Animal Production
(College of Abouraihan – University of Tehran)
Vol. 19 ■ No. 1 ■ Spring 2017

Comparison of parametric and resampling methods in genetic evaluation of quantitative traits with different genetic structure

Manoocher Moradi[1] , Rostam Abdollahi-Arpanahi[2], Behzad Hemati[3] , Abolghasem Lavvaf[3]*

1. M.Sc. student, Department of Animal Science, Faculty of Agriculture and Natural Resources , Karaj Branch, Islamic Azad University, Karaj, Iran
2. Assistant Professor, Department of Animal and Poultry Science, College of Aburaihan, University of Tehran, Tehran, Iran.
3. Associate Professor, Department of Animal Science, Faculty of Agriculture and Natural Resources, Karaj Branch, Islamic Azad University, Karaj, Iran

Received: May 26, 2016 **Accepted: September 20, 2016**

Abstract

The objective of this study was to compare three parametric (GBLUP, BayesB and RKHS) and two resampling (Bagging GBLUP and Random Forest) statistical methods in genomic prediction of traits with different genetic architecture. A genome consisting of three chromosomes, 1 Morgan each, was simulated on which 5000 SNPs and 50, 100 and 200 QTLs were distributed. The substitutions effects of QTLs were modeled with normal, gamma and uniform distributions with a level of heritability equal to 0.30. The predictive performance of statistical models was evaluated using the correlation between predicted and true breeding values as well as the regression of predicted values on true breeding values. In the target population, Random Forest resulted in overestimation of estimated regression coefficients while GBLUP, BayesB and RKHS led to an underestimation of regression coefficients of true breeding values on predicted breeding values. In exception of Bagging GBLUP, the performance of all statistical methods was the same in three gene effect distributions. However, the performance of GBLUP and BayesB was better than other statistical methods. A reason for this superiority could be the additive architecture of simulated traits. In conclusion, GBLUP and BayesB were superior over resampling methods in genomic predictions.

Keywords: Bagging, Genetic architecture, Genomic evaluation, Machine learning, Random forest**,** RKHS

Email: r.abdollahi@ut.ac.ir